PRIFYSGOL ABERYSTWYTH UNIVERSITY

Aberystwyth University

Are More Features Better?

Jensen, Richard; Shen, Qiang

Published in:
IEEE Transactions on Fuzzy Systems

DOI:
10.1109/TFUZZ.2009.2026639

Publication date:
2009

Citation for published version (APA):
Jensen, R., & Shen, Q. (2009). Are More Features Better? A Response to Attributes Reduction Using Fuzzy Rough Sets. *IEEE Transactions on Fuzzy Systems*, *17*(6), 1456-1458. https://doi.org/10.1109/TFUZZ.2009.2026639

General rights
Copyright and moral rights for the publications made accessible in the Aberystwyth Research Portal (the Institutional Repository) are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

• Users may download and print one copy of any publication from the Aberystwyth Research Portal for the purpose of private study or research.
• You may not further distribute the material or use it for any profit-making activity or commercial gain
• You may freely distribute the URL identifying the publication in the Aberystwyth Research Portal

Take down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

tel: +44 1970 62 2400
email: is@aber.ac.uk

Download date: 04. May. 2023

Letters

Are More Features Better? A Response to *Attributes Reduction Using Fuzzy Rough Sets*

Richard Jensen and Qiang Shen

***Abstract*—A recent TRANSACTIONS ON FUZZY SYSTEMS paper proposing a new fuzzy-rough feature selector (FRFS) has claimed that the more attributes remain in datasets, the better the approximations and hence resulting models. [Tsang *et al.*, *IEEE Trans. Fuzzy Syst.*, vol. 16, no. 5, pp. 1130–1141]. This claim has been used as a primary criticism of the original FRFS method [Jensen and Shen, *IEEE Trans. Fuzzy Syst.*, vol. 15, no. 1, pp. 73–89, Feb. 2007]. Although, in certain applications, it may be necessary to consider as many features as possible, the claim is contrary to the motivation behind feature selection concerning the curse of dimensionality, the presence of redundant and irrelevant features, and the large amount of literature documenting observed improvements in modeling techniques following data reduction. This letter discusses this issue, as well as two other issues raised by Tsang *et al.* [*IEEE Trans. Fuzzy Syst.*, vol. 16, no. 5, pp. 1130–1141, Oct. 2008] regarding the original algorithm.**

***Index Terms*—Dimensionality reduction, feature selection (FS), fuzzy-rough sets.**

I. INTRODUCTION

FEATURE selection (FS) [3], [15] addresses the problem of selecting those input features that are most predictive of a given outcome, which is a problem encountered in many areas of computational intelligence [12]. Unlike other dimensionality-reduction methods, FSs preserve the original meaning of the features after reduction. This has found application in tasks that involve datasets containing huge numbers of features (in the order of tens of thousands) that, for some learning algorithms, might be impossible to process further. Recent examples include text processing and Web content classification [8]. There are often many features involved and, combinatorially, large numbers of feature combinations, from which to select.

Fuzzy-rough FS (FRFS) [11] provides a means by which discrete or real-valued noisy data (or a mixture of both) can be effectively reduced without the need for user-supplied information. Additionally, this technique can be applied to data with continuous or nominal decision attributes and, as such, can be applied to regression, as well as classification datasets. The only additional information required is in the form of fuzzy partitions for each feature that can be automatically derived from the data. FRFS has been shown to be a highly useful technique in reducing data dimensionality. However, there are several problems with the approach from theoretical and practical viewpoints that motivate further developments in this area, including work reported in [2], [7], [13], and [24].

In particular, recent research in [24] presented an alternative and significant approach to FS with fuzzy-rough sets. However, a controversial claim was made in this paper that deserves a fuller analysis and discussion—that retaining more attributes in datasets will lead to better approximations. This can be the case for perfect, entirely consistent, and noise-free data, with all features being independent. Nevertheless, no reduction is theoretically possible in such situations without losing information. Of course, it may be necessary to consider as many features as possible for certain application problems. Yet, the viewpoint of "the more the features, the better the approximations" is, in general, a particularly strong claim to make, given the abundance of research in the FS area that seeks to improve models and approximations through the elimination of irrelevant, redundant, and noisy features. This issue and two further important ones, as raised in [24] regarding fuzzy equivalence class construction and computational cost of the original work in [11], are addressed in the next sections.

II. MORE ATTRIBUTES OFFER BETTER APPROXIMATIONS?

One of the motivating factors for the research presented in [24] seemed to have been the observation that the original FRFS dependency measure in [11] is nonmonotonic. As a result of nonmonotonicity, a feature subset P_1 with a smaller cardinality than a feature subset P_2 may, in fact, have a higher fuzzy-rough dependency degree and is, therefore, more desirable from the perspective of FS methods.

It is this situation that the view expressed in [24] regarded it as being "unreasonable." However, the general claim that the use of fewer features will result in poorer approximation ability is problematic for a number of reasons, including the following.

A. Curse of Dimensionality

The curse of dimensionality [1], when considering learning tasks, describes the problem facing learning methods where an increasing number of features require an exponentially increasing number of training objects. In other words, as the dimensionality increases, objects in the training data become too sparse to train learning algorithms effectively. In situations where there are a large number of features and relatively few objects

Manuscript received November 7, 2008; revised May 29, 2009; accepted May 29, 2009. First published July 7, 2009; current version published December 3, 2009.

The authors are with the Department of Computer Science, University of Wales, Aberystwyth SY23 3DB, U.K. (e-mail: rkj@aber.ac.uk; qqs@aber.ac.uk).

Digital Object Identifier 10.1109/TFUZZ.2009.2026639

1063-6706/$26.00 © 2009 IEEE

Authorized licensed use limited to: IEEE Transactions on SMC Associate Editors. Downloaded on December 5, 2009 at 11:16 from IEEE Xplore. Restrictions apply.

(e.g., the type of data often obtained in bioinformatics experimentation), dimensionality reduction is essential. Attempting to approximate and model concepts using the full feature set is futile given the data sparsity; therefore, feature subsets of large cardinality may be considered to be much less informative than those of a smaller cardinality. Obviously, there is a point where feature subsets must be of sufficient size in order to adequately describe the concepts, but this is usually at least one order of magnitude smaller than the full feature set.

From the traditional rough set perspective, the monotonicity of the dependency measure means that the inclusion of more features will lead to increasingly smaller knowledge granules, which will allow concepts to be more easily approximated. However, this does not necessarily mean that the approximations are *better* when using larger numbers of features but only means that they are *easier* to construct. In the recent developments reported in [13], as well as for many crisp rough set feature selectors [6], the FS process terminates when a subset has been found that preserves the positive region of the entire feature set. In this case, the addition of any remaining features to the subset will not change the positive region and, therefore, will not improve the underlying concept approximations. Indeed, it may be detrimental to the quality of the approximations to include more features.

B. Feature Redundancy and Relevancy

There are at least two feature qualities that must be considered by the FS methods: relevancy and redundancy [3], [14]. A feature is said to be relevant if it is predictive of the decision feature(s); otherwise, it is irrelevant. A feature is considered to be redundant if it is highly correlated with other features. An informative feature is one that is highly correlated with the decision concept(s) but is highly uncorrelated with other features (although low correlation does not mean absence of relationship). Similarly, subsets of features should exhibit these properties of relevancy and nonredundancy if they are to be useful in an efficient manner. It may also be the case that a dataset contains noisy or misleading features, as is often the case when extracting information from the real world.

The removal of such features that may be noisy, misleading, redundant, or irrelevant must surely *improve* the quality of approximations constructed via machine-learning methods. In this way, it may well be the case that the quality of the full set or a superset of features is less than that of a subset with such undesirable features removed. For the purposes of FS, a nonmonotonic measure of subset goodness may, in fact, be more sensible.

C. Experimental Verification

There has been considerable research into the area of FS, with many reported results of statistically significant improvement in approximation quality after selection has been performed. In particular, a large volume of published results in the relevant literature have demonstrated that smaller subsets of selected features can lead to much-improved modeling accuracy. For recent work, see [5], [9], [16]–[18], and [25]. Much evidence also exists in work adopting a rough or fuzzy-rough set approach to FS. For example, in the Web content categorization domain, with data possessing thousands of features, significant data reduction can be achieved while maintaining or even improving classification performance [8]. This is also witnessed in work about classification of medical images [19] and gene expressions [20]. A similar trend was observed for the application of FRFS to systems monitoring [22] and, more recently, to algae population estimation [23] and forensic glass fragment classification [10].

III. Fuzzy Equivalence Class Construction

In rough set theory, concept approximations are constructed via the manipulation of equivalence classes. The extension of such classes is, therefore, fundamental to the success of any approach that attempts to extend rough set methods in this way, for example, tolerance-based [10], [16] and fuzzy-rough methods [12], [24] for FS. In particular, for fuzzy extensions, the construction of fuzzy equivalence classes is an important issue.

It is stated in [24] that "the Cartesian product of $\mathbb{U}/\text{IND}(\{a\})$ and $\mathbb{U}/\text{IND}(\{b\})$ may not be a collection of fuzzy equivalence classes of a fuzzy similarity relation, even this statement holds when a and b are two crisp equivalence relations." This may be misleading for fuzzy equivalence class construction. In the original FRFS (and crisp rough set attribute reduction [21]), however, a and b are not fuzzy similarity relations but are features. $\text{IND}(\{a\})$ is the indiscernibility relation induced by feature a, and $\mathbb{U}/\text{IND}(\{a\})$ is the resulting partitioning of the universe of discourse by the relation. In the fuzzy case, the family of normal fuzzy sets produced by a fuzzy partitioning of the universe of discourse play the role of fuzzy equivalence classes [4]. The n-ary Cartesian product of these fuzzy equivalence classes is used when more than one feature is under consideration. In the crisp case, the n-ary Cartesian product of crisp equivalence classes is used and will also always produce a collection of equivalence classes.

IV. Computational Cost

The experimentation section of [24] investigates the application of several FRFS algorithms to the wine dataset. In this section, it was reported that the original FRFS algorithm did not terminate due to its high computational cost. This is surprising, particularly given the past successful application of FRFS to data exhibiting a much higher dimensionality [8].

It is agreed that the complexity of calculating the n-ary Cartesian product of fuzzy equivalence classes can become prohibitively high for large feature subsets. This observation was also made in [2]. If the number of fuzzy sets per attribute is n, $n^{|R|}$ equivalence classes must be considered per feature for feature subset R of cardinality $|R|$. However, there are a number of optimizations that largely alleviate this problem, as given in [2] and [11], using the properties of the fuzzy connectives employed in the reduct search algorithm, as well as data structures, to avoid unnecessary computations.

Nevertheless, this issue is now largely redundant following recent developments in FRFS, particularly the work in [13]. In

Authorized licensed use limited to: IEEE Transactions on SMC Associate Editors. Downloaded on December 5, 2009 at 11:16 from IEEE Xplore. Restrictions apply.

this latter research, a general methodology for FRFS is presented based on fuzzy tolerance relations. This approach guarantees dependency function monotonicity, avoids the n-ary Cartesian product calculations, and improves both the time and space complexity of the underlying search algorithms.

V. Conclusion

This letter has addressed three issues raised in [24] regarding the original FRFS method, concerning the nonmonotonicity of the subset evaluation measure, the construction of fuzzy equivalence classes, and the computational complexity. In particular, the claim that retaining more attributes in datasets will lead to better approximations has been argued against, with respect to the curse of dimensionality, feature redundancy and relevancy, and the wealth of experimental results from the FS community. Indeed, it may be the case that nonmonotonicity is a desirable property of subset evaluation measures when considering these factors. Obviously, from a practical viewpoint, the utility of FS methods is dependent on both the data under consideration and the intended purpose of the data analysis. Under certain circumstances, it is necessary to consider as many features as possible, and in others, a minimal subset of features satisfying some predefined criteria is desired. From a theoretical viewpoint, it is only the case when the original data are fully consistent and noise-free, and further, all of its features are independent; retaining more features will result in better approximations. Yet, in such cases, no feature reduction is possible without losing information.

Acknowledgment

The authors would like to thank the referees for their invaluable and insightful comments that have helped to shape this letter and further discussions in this important area.

References

[1] R. E. Bellman, *Adaptive Control Processes*. Princeton, NJ: Princeton Univ. Press, 1961.

[2] R. B. Bhatt and M. Gopal, "On the compact computational domain of fuzzy-rough sets," *Pattern Recognit. Lett.*, vol. 26, no. 11, pp. 1632–1640, 2005.

[3] M. Dash and H. Liu, "Feature selection for classification," *Intell. Data Anal.*, vol. 1, no. 3, pp. 131–156, 1997.

[4] D. Dubois and H. Prade, "Putting rough sets and fuzzy sets together," in *Intelligent Decision Support*. London, U.K.: Kluwer, 1992, pp. 203–232.

[5] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *J. Mach. Learn. Res.*, vol. 3, pp. 1157–1182, 2003.

[6] A. E. Hassanien, Z. Suraj, D. Ślęzak, and P. Lingras, Eds., *Rough Computing: Theories, Technologies and Applications*. Hershey, NY: Inf. Sci. Ref., 2008.

[7] Q. Hu, D. Yu, and Z. Xie, "Information-preserving hybrid data reduction based on fuzzy-rough techniques," *Pattern Recognit. Lett.*, vol. 27, no. 5, pp. 414–423, 2006.

[8] R. Jensen and Q. Shen, "Fuzzy-rough attribute reduction with application to web categorization," *Fuzzy Sets Syst.*, vol. 141, no. 3, pp. 469–485, 2004.

[9] R. Jensen and Q. Shen, "Semantics-preserving dimensionality reduction: Rough and fuzzy-rough based approaches," *IEEE Trans. Knowl. Data Eng.*, vol. 16, no. 12, pp. 1457–1471, Dec. 2004.

[10] R. Jensen and Q. Shen, "Tolerance-based and Fuzzy-rough feature selection," in *Proc. 16th Int. Conf. Fuzzy Syst.*, 2007, pp. 877–882.

[11] R. Jensen and Q. Shen, "Fuzzy-rough sets assisted attribute selection," *IEEE Trans. Fuzzy Syst.*, vol. 15, no. 1, pp. 73–89, Feb. 2007.

[12] R. Jensen and Q. Shen, *Computational Intelligence and Feature Selection: Rough and Fuzzy Approaches*. Piscataway, NJ: Wiley–IEEE Press, 2008.

[13] R. Jensen and Q. Shen, "New approaches to fuzzy-rough feature selection," *IEEE Trans. Fuzzy Syst.*, to be published.

[14] P. Langley, "Selection of relevant features in machine learning," in *Proc. AAAI Fall Symp. Relevance*, 1994, pp. 1–5.

[15] H. Liu and H. Motoda, *Feature Selection for Knowledge Discovery and Data Mining*. Norwell, MA: Kluwer, 1998.

[16] N. M. Parthalain and Q. Shen, "Exploring the boundary region of tolerance rough sets for feature selection," *Pattern Recognit.*, vol. 42, no. 5, pp. 655–667, 2009.

[17] N. Mac Parthalain, Q. Shen, and R. Jensen, "A distance measure approach to exploring the rough set boundary region for attribute reduction," *IEEE Trans. Knowl. Data Eng.*, to be published.

[18] Y. Saeys, I. Inza, and P. Larrañaga, "A review of feature selection techniques in bioinformatics," *Bioinformatics*, vol. 23, no. 19, pp. 2507–2517, 2007.

[19] C. Shang and Q. Shen, "Rough feature selection for neural network based image classification," *Int. J. Image Graph.*, vol. 2, no. 4, pp. 541–555, 2002.

[20] C. Shang and Q. Shen, "Aiding classification of gene expression data with feature selection: A comparative study," *Comput. Intell. Res.*, vol. 1, no. 1, pp. 68–76, 2006.

[21] Q. Shen and A. Chouchoulas, "A rough-fuzzy approach for generating classification rules," *Pattern Recognit.*, vol. 35, no. 11, pp. 2425–2438, 2002.

[22] Q. Shen and R. Jensen, "Selecting informative features with fuzzy-rough sets and its application for complex systems monitoring," *Pattern Recognit.*, vol. 37, no. 7, pp. 1351–1363, 2004.

[23] Q. Shen and R. Jensen, "Approximation-based feature selection and application for algae population estimation," *Appl. Intell.*, vol. 28, no. 2, pp. 167–181, 2008.

[24] E. C. C. Tsang, D. Chen, D. S. Yeung, X.-Z. Wang, and J. W. T. Lee, "Attributes reduction using fuzzy rough sets," *IEEE Trans. Fuzzy Syst.*, vol. 16, no. 5, pp. 1130–1141, Oct. 2008.

[25] L. Yu and H. Liu, "Efficient feature selection via analysis of relevance and redundancy," *J. Mach. Learn. Res.*, vol. 5, pp. 1205–1224, 2004.

Authorized licensed use limited to: IEEE Transactions on SMC Associate Editors. Downloaded on December 5, 2009 at 11:16 from IEEE Xplore. Restrictions apply.